"Express Mail" Mailing Label No. **E960827938US**

PATENT APPLICATION
ATTORNEY DOCKET NO. TEK03-1002

METHOD AND APPARATUS FOR DYNAMICALLY ALLOCATING UPSTREAM BANDWIDTH IN PASSIVE OPTICAL NETWORKS

Inventors: John F. Sisto and Edward W. Boyd

BACKGROUND

Field of the Invention

[0001] The present invention relates to the design of passive optical networks. More specifically, the present invention relates to a method and apparatus for dynamically allocating upstream bandwidth in a passive optical network.

Related Art

[0002] In order to keep pace with increasing Internet traffic, optical fibers and associated optical transmission equipment have been widely deployed to substantially increase the capacity of backbone networks. However, this increase

in the capacity of backbone networks has not been matched by a corresponding increase in the capacity of access networks. Even with broadband solutions, such as digital subscriber line (DSL) and cable modem (CM), the limited bandwidth offered by current access networks creates a severe bottleneck in delivering high bandwidth to end users.

[0003] Among different technologies, Ethernet passive optical networks (EPONs) appear to be the best candidate for next-generation access networks. EPONs combine the ubiquitous Ethernet technology with inexpensive passive optics. Therefore, they offer the simplicity and scalability of Ethernet, and the cost-efficiency and high capacity of passive optics. In particular, due to the high bandwidth of optical fibers, EPONs are capable of accommodating broadband voice, data, and video traffic simultaneously. Such integrated service is difficult to provide with DSL or CM technology. Furthermore, EPONs are more suitable for Internet Protocol (IP) traffic, since Ethernet frames can directly encapsulate native IP packets with different sizes, whereas ATM passive optical networks (APONs) use fixed-size ATM cells and consequently require packet fragmentation and reassembly.

[0004] Typically, EPONs are used in the "first mile" of the network, which provides connectivity between the service provider's central offices and business or residential subscribers. Logically, the first mile is a point-to-multipoint network, with a central office servicing a number of subscribers. A tree topology can be used in an EPON, wherein one fiber couples the central office to a passive optical splitter, which divides and distributes downstream optical signals to subscribers and combines upstream optical signals from subscribers (see FIG. 1).

[0005] Transmissions within an EPON are typically performed between an optical line terminal (OLT) and optical networks units (ONUs) (see FIG. 2). The

OLT generally resides in the central office and couples the optical access network to the metro backbone, which is typically an external network belonging to an ISP or a local exchange carrier. The ONU can be located either at the curb or at an end-user location, and can provide broadband voice, data, and video services.

[0006] Communications within an EPON can be divided into upstream traffic (from ONUs to OLT) and downstream traffic (from OLT to ONUs). Because of the broadcast nature of Ethernet, the downstream traffic can be delivered with considerable simplicity in an EPON: packets are broadcast by the OLT and extracted by their destination ONU based on their media access control (MAC) addresses. However, in the upstream direction, the ONUs need to share the channel capacity and resources. Moreover, the burstiness of network traffic and the requirement of different service level agreements (SLAs) make the upstream bandwidth allocation a challenging problem.

[0007] Hence, what is needed is a method and apparatus for dynamically allocating upstream bandwidth in an EPON, which is fair, efficient, and responsive, and which accommodates bursty traffic while satisfying SLAs.

SUMMARY

[0008] One embodiment of the present invention provides a system that facilitates dynamic allocation of upstream bandwidth in a passive optical network which includes a central node and at least one remote node. Each remote node is coupled to at least one logical entity, which corresponds to a device or a user, that transmits upstream data to the central node and receives downstream data from the central node. The central node is coupled to an external network outside of the passive optical network through a shared out-going uplink.

[0009] During operation, the system receives a request from a remote node for a grant to transmit upstream data from a logical entity associated with the

remote node to the central node, wherein the size of the data to be transmitted does not exceed a transmission threshold assigned to that logical entity, and a logical entity may not request more than what is allowed by the corresponding transmission threshold. If the request satisfies a bandwidth allocation policy, the system issues a grant to the remote node to transmit upstream data. In response to the grant, the system receives upstream data from the remote node and places the received upstream data in a receiver buffer within the central node. This receiver buffer includes a number of FIFO queues, each of which buffers upstream data received from an associated logical entity. Next, the system retrieves and transmits data stored in the receiver buffer to the out-going uplink according to a set of SLAs.

[0010] In a variation of this embodiment, satisfying the bandwidth allocation policy requires that there be sufficient available space in the receiver buffer to accommodate the upstream data to be transmitted, and that the logical entity from which upstream data transmission is requested is scheduled to transmit data next.

[0011] In a further variation, all the logical entities within the passive optical network are scheduled to transmit upstream data in a hierarchical round-robin scheme by performing the following operations:

[0012] (1) grouping logical entities with the highest priority to form a top-priority level;

[0013] (2) allowing each logical entity in the top-priority level to transmit upstream data in a round-robin fashion by assigning a slot to each logical entity in the top-priority level;

[0014] (3) within the top-priority level, reserving one slot for lower-priority traffic;

[0015] (4) grouping logical entities with the second-highest priority to form a second-priority level;

[0016] (5) allowing each logical entity in the second-priority level to transmit data by assigning the reserved slot within the top-priority level to each logical entity in the second-priority level in a round-robin fashion;

[0017] (6) within the second-priority level, reserving one slot for lower-priority traffic; and

[0018] (7) repeating operations similar to operations (4) – (6) for logical entities with lower priorities until every logical entity is assigned a slot for transmitting upstream data according to its priority.

[0019] In a variation of this embodiment, the transmission threshold assigned to a logical entity within a priority level is determined by considering the maximum allowable delay for that priority level, data speed of the shared out-going uplink, the logical entity's SLA, and the total number of logical entities within that priority level.

[0020] In a variation of this embodiment, the system keeps a record of outstanding data for each logical entity, wherein outstanding data is upstream data for which a grant for transmission has been issued by the central node, but which has not been received by the central node. To calculate available space in the receiver buffer, the system subtracts the size of outstanding data from the unfilled space of the corresponding FIFO queue. After a period of time following issuance of a grant for transmitting a piece of data, the data is due to arrive at the system. The system accordingly removes the information pertinent to the piece of data from the record of outstanding data for the corresponding logical entity, which is done regardless of whether the piece of data has actually been received by the central node.

[0021] In a variation of this embodiment, the system retrieves and transmits data stored in each FIFO queue within the receiver buffer in a hierarchical round-robin scheme in accordance with each logical entity's SLA.

[0022] In a variation of this embodiment, each remote node includes a number of queues, each of which is associated with a logical entity and stores upstream data from the device or user associated with that logical entity.

[0023] In a further variation, the request from the remote node reports the state of a queue within that remote node associated with a logical entity, and the request piggybacks on an upstream data transmission.

[0024] In a further variation, if a FIFO queue within the receiver buffer in the central node is full, the system pauses the issuing of grants to the corresponding logical entity, thereby causing the queue associated with that logical entity within a remote node to become full. This causes the remote node to generate a flow-control message to the corresponding device or user to slow down or pause the upstream data transmission from that device or user.

[0025] In a variation of this embodiment, a remote node tracks the amount of time between the grants to transmit upstream data for each logical entity associated with the remote node. If the amount of time between grants exceeds a certain interval, the remote node sets an alarm and sends a message to the central node via an Operation, Administration and Maintenance (OAM) frame, whereby upon receiving the message, the central node is allowed to reset a record associated with the corresponding logical entity.

[0026] In a variation of this embodiment, the central node periodically sends out polls to the remote nodes to see if a logical entity has any data to send. The polling frequency for a corresponding logical entity reflects the SLA of the logical entity. If a non-poll grant has been previously sent to a logical entity, the

subsequent poll to that logical entity is sent at a time after the non-poll grant, the time being calculated in accordance to the corresponding polling frequency.

[0027] In a further variation, a remote node tracks the amount of elapsed time between non-poll grants for each logical entity associated with the remote node. If the elapsed time between non-poll grants for a logical entity exceeds a certain interval, the remote node sets an alarm. If the alarm is set and the remote node has data to send from the corresponding logical entity, the remote node sends a message to the central node via an OAM frame denoting an error condition, which instructs the central node that the logical entity is in an error state. Upon receiving the message, the central node is allowed to reset or modify a record associated with the logical entity.

BRIEF DESCRIPTION OF THE FIGURES

[0028] FIG. 1 illustrates a passive optical network wherein a central office and a number of subscribers form a tree topology through optical fibers and a passive optical splitter (prior art).

[0029] FIG. 2 illustrates a passive optical network including an OLT and ONUs (prior art).

[0030] FIG. 3 illustrates the architecture of an OLT that facilitates dynamic upstream bandwidth allocation in accordance with an embodiment of the present invention.

[0031] FIG. 4 presents a flow chart illustrating the dynamic upstream bandwidth allocation process in accordance with an embodiment of the present invention.

[0032] FIG. 5 illustrates a flow-control mechanism within an OLT that facilitates dynamic upstream bandwidth allocation in accordance with an embodiment of the present invention.

[0033] FIG. 6 illustrates a hierarchical round-robin scheduling scheme with transmission thresholds in accordance with an embodiment of the present invention.

[0034] FIG. 7 illustrates a time-out mechanism for outstanding data that provides fault tolerance in accordance with an embodiment of the present invention.

DETAILED DESCRIPTION

[0035] The following description is presented to enable any person skilled in the art to make and use the invention, and is provided in the context of a particular application and its requirements. Various modifications to the disclosed embodiments will be readily apparent to those skilled in the art, and the general principles defined herein may be applied to other embodiments and applications without departing from the spirit and scope of the present invention. Thus, the present invention is not intended to be limited to the embodiments shown, but is to be accorded the widest scope consistent with the principles and features disclosed herein.

[0036] The data structures and code described in this detailed description are typically stored on a computer readable storage medium, which may be any device or medium that can store code and/or data for use by a computer system. This includes, but is not limited to, application specific integrated circuits (ASICs), field-programmable gate arrays (FPGAs), semiconductor memories, magnetic and optical storage devices such as disk drives, magnetic tape, CDs (compact discs) and DVDs (digital versatile discs or digital video discs), and computer instruction signals embodied in a transmission medium (with or without a carrier wave upon which the signals are modulated). For example, the

transmission medium may include a communications network, such as the Internet.

Passive Optical Network Topology

[0037] FIG. 1 illustrates a passive optical network, wherein a central office and a number of subscribers form a tree topology through optical fibers and a passive optical splitter. As shown in FIG. 1, a number of subscribers are coupled to a central office 101 through optical fibers and a passive optical splitter 102. Passive optical splitter 102 can be placed in the vicinity of end-user locations, so that the initial fiber deployment cost is minimized. The central office is coupled to an external network, such as a metropolitan area network operated by an ISP.

[0038] FIG.2 illustrates a passive optical network including an OLT and ONUs. OLT 201 is coupled with ONUs 202, 203, and 204 through optical fibers and a passive optical splitter. An ONU can accommodate a number of networked devices, such as personal computers, telephones, video equipment, network servers, etc. Note that a networked device can identify itself by using a Logical Link ID (LLID), as defined in the IEEE 802.3 standard.

Dynamic Bandwidth Allocation Mechanism

[0039] FIG. 3 illustrates the architecture of an OLT that facilitates dynamic upstream bandwidth allocation in accordance with an embodiment of the present invention. In this example, an OLT 320 accepts requests and upstream data traffic from ONUs 301 and 302. Each ONU maintains a number of queues, for example queues 311, 312, and 313, each of which stores upstream data from an LLID corresponding to a device or a user that couples to that ONU. Note that upstream data from an LLID is carried in data frames (e.g., Ethernet frames),

which have variable sizes. During transmission these data frames are removed from their respective queue. An LLID requests a grant, to transmit upstream data, via a report message. The report message indicates the amount of data in the LLID's corresponding queue(s). Typically, these request messages can piggyback on an upstream data transmission.

[0040] Within OLT 310, a dynamic bandwidth allocation (DBA) scheduler 303 receives the report messages from ONUs. OLT 310 also includes a FIFO queue controller (FCT) 305, which contains a number of FIFO queues (321, 322, 323, 324, and 325) that are associated with different LLIDs. Upstream data from each LLID is temporarily stored in these FIFO queues before being transmitted to the external ISP network through a shared uplink 330. The state of these FIFO queues is monitored and stored in a queue length table 304.

[0041] After receiving a request from an LLID, DBA scheduler 303 determines whether a grant to transmit can be sent to the requesting LLID based on two considerations. First, whether there is sufficient available space in the FIFO queue corresponding to the requesting LLID, according queue length table 304. Second, whether the requesting LLID is the next in turn to transmit data as scheduled. (Note that proper scheduling of LLIDs for upstream data transmission is necessary to guarantee fair and efficient bandwidth allocation among all the LLIDs.) When both conditions are met, the DBA scheduler issues a grant to the requesting LLID. The grant allocates an upstream transmission time slot to the LLID.

[0042] Note that outstanding data for each LLID can be taken into account in the calculation of available space in the FIFO queues. Outstanding data is the "in-flight" data for which a grant for transmission has been given, but which has not been received by OLT 320. Records of outstanding data are stored in data structure 309. When calculating available space in a FIFO queue, DBA scheduler

303 subtracts the amount of outstanding data of the requesting LLID from the available physical space in the corresponding FIFO queue, and uses the result as the actual available space for future data transmission.

[0043] With regard to scheduling upstream transmission, one possible scheme is the hierarchical round-robin scheme, which can be used to fairly and efficiently allocate bandwidth among all LLIDs. Another possible scheduling scheme is strict priority scheduling. However, because SLAs usually place constraints on parameters such as average bit rate, maximum delay, etc., a transmission threshold (the maximum amount of data in each transmission) may be set for every LLID in the hierarchical round-robin scheme. A more detailed discussion of this scheme appears in the discussion related to FIG. 5 below.

[0044] OLT 320 further includes a bandwidth shaper 307, which retrieves data stored in the FIFO queues within FCT 305 and transmits the retrieved data to shared uplink 330. Bandwidth shaper 307 ensures that the data stored in FCT 305 is served in accordance with the priority classification and SLA pertinent to each LLID, which is stored in data structure 306. Like the scheduling mechanism within DBA scheduler 303, the scheduling mechanism within bandwidth shaper 307 is desired to be fair and efficient, and therefore can also use the hierarchical round-robin scheduling scheme.

[0045] FIG. 4 presents a flow chart illustrating the dynamic upstream bandwidth allocation process in accordance with an embodiment of the present invention. The system starts by receiving a report message from an LLID at the DBA scheduler 303 (step 401). DBA scheduler 303 then determines if there is sufficient space in the FIFO queue within FCT 305 for this LLID (taking into account the outstanding data) (step 402). If there is not sufficient space, DBA scheduler temporarily holds the grant for the requesting LLID until sufficient

space becomes available in the FIFO queue. Meanwhile, the system can receive and process requests from other LLIDs by returning to step 401.

[0046] If there is sufficient space in the FIFO queue within FCT 305, DBA scheduler 303 further determines if the requesting LLID is scheduled to transmit data next (step 403). If not, DBA scheduler 303 will temporarily hold the grant until the requesting LLID is the next to transmit. Meanwhile, the system can receive and process requests from other LLIDs by returning to step 401.

[0047] If it is the requesting LLID's turn to transmit, DBA scheduler generates a grant and sends it to the requesting LLID (step 404). The system then returns to step 401 and continues to receive and process subsequent requests.

Flow-control Mechanism

[0048] FIG. 5 illustrates a flow-control mechanism within an OLT that facilitates dynamic upstream bandwidth allocation in accordance with an embodiment of the present invention. In this example, when FIFO queue 323 is filled, DBA scheduler 303 stops granting transmission from LLID #3, thereby causing queue 313 to fill. ONU 302 can then generate a flow-control message in accordance with the IEEE 802.3x standard to the corresponding device or user to slow down, or pause, further upstream data transmission.

Hierarchical Round-robin Scheduling with Transmission Thresholds

[0049] FIG. 6 illustrates a hierarchical round-robin scheduling scheme with transmission thresholds in accordance with an embodiment of the present invention. This hierarchical round-robin scheduling is performed as follows:

[0050] First, group all LLIDs with the highest priority (priority 0). Within priority 0, assign each LLID a transmission slot in accordance to an amount of data burst the LLID is allowed to transmit upstream. The LLID is

provisioned to not report a value greater than this amount. Although the aggregate of all report messages in a report frame may exceed this threshold, the amount of data implied in each individual message cannot exceed this burst size. The slot size provisioned for each LLID is determined such that all the LLIDs may be serviced within a fixed delay bounds. For example, if the delay bounds for priority 0 is one ms, and shared uplink 330's data speed is 1 Gb/s, then the total duration of priority 0 may not exceed 1000 Kb. Therefore, the aggregate slot size of priority 0 LLIDs would sum up to less than or equal to 1000 Kb.

[0051] Within priority 0, one slot is allocated for lower priority traffic. This slot is denoted as the drop-down slot. All lower-priority traffic is allowed to transmit within this reserved slot.

[0052] Next, group all of the LLIDs with the second highest priority (priority 1). Within priority 1, assign each LLID a transmission slot according to the maximum burst the LLID may transmit upstream. The LLID will be configured such that it will observe this maximum burst size when reporting. A slot in priority 1 is allowed to transmit inside the slot reserved for lower-priority traffic (the drop-down slot) within priority 0. Since a priority 1 LLID may only transmit when priority 0 is transmitting its drop-down slot, the delay of the queuing delay of priority 1 LLIDs is typically many times of the queuing delay of priority 0 LLIDs.

[0053] Within priority 1, there is similarly one slot reserved for lower-priority traffic.

[0054] As shown in FIG. 6, one can repeat steps similar to the above, and construct an entire hierarchy to accommodate all the LLIDs. Note that the transmission thresholds of LLIDs within a given priority level is based on the bandwidth and maximum allowable delay negotiated in the corresponding SLA.

Fault Tolerance

[0055] FIG. 7 illustrates a time-out mechanism for outstanding data that provides fault tolerance in accordance with an embodiment of the present invention. During operation, it is possible that a grant message 731 is lost on its way from OLT 720 to ONU 610, for example due to a bit error. As a result, the subsequent grant messages received by ONU 710 for the same LLID will grant transmission sizes that are inconsistent with the amount of data available for upstream transmission. This may manifest itself by the ONU receiving a grant that is not a frame boundary. Once ONU 710 detects this inconsistency, it will start sending special report messages to OLT 720, requesting a transmission size of 0 Kb. Meanwhile, OLT 720 keeps track of when a piece of upstream data associated with a grant is due to arrive. Whether or not this piece of data physically arrives for the grant, the OLT removes the information corresponding to the outstanding data for the grant.

[0056] After sending the special report messages (with request of 0 K) for a period of time, ONU 710 resumes sending normal request messages. By this time the lost grant message, and its residual effects, would have timed out in OLT 720 and normal operation resumes.

[0057] It is possible for an ONU to track the amount of time between grants. If the amount of time between grants exceeds a certain interval, ONU 710 sets an alarm and sends a message to OLT 720 via an OAM frame. This can be done via an LLID on the ONU that is reserved for processor traffic. This message will instruct OLT 720 that an LLID is not being granted. One way for OLT 720 to deal with this situation is to reset the LLID entry in the DBA and bandwidth shaper tables.

[0058] In another scenario, OLT 720 periodically sends out polls to ONUs to see if an LLID has any data to send. Polls are grants for 64 bytes of data that

have a forced-report flag asserted. The only upstream data transmitted as a response to a poll is a single report frame. The polling frequency reflects the SLA of an LLID. For example, the polls for priority 0 LLIDs are sent every 1 ms. If a grant previously occurred, the subsequent poll will be sent at 1 ms after that grant being sent.

[0059] Correspondingly, a non-poll grant is a grant that allows transmission of more than just a single report frame. An ONU tracks the amount of time elapsed between non-poll grants for each LLID. If this time exceeds a certain interval, the ONU sets an alarm. If the alarm is set, and the ONU has data to send, the ONU will send a message to the OLT, via an OAM frame, denoting the error condition. This will instruct the OLT that an LLID is in an error state. One way for the OLT to deal with this situation is to reset or modify the LLID entry in the DBA and bandwidth scheduler tables.

[0060] The foregoing descriptions of embodiments of the present invention have been presented for purposes of illustration and description only. They are not intended to be exhaustive or to limit the present invention to the forms disclosed. Accordingly, many modifications and variations will be apparent to practitioners skilled in the art. Additionally, the above disclosure is not intended to limit the present invention. The scope of the present invention is defined by the appended claims.